Advanced Regression: Random effects and hierarchical models II

Garyfallos Konstantinoudis

Feb 20, 2024

4. Random effects

1. Random effect model with random intercept

$$y_i = (\alpha_0 + u_k) + \beta x_i + \epsilon_i,$$

   where $u_k \sim N(0, \sigma_u^2)$.

2. Random effects model on both, the intercept and the slope

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \epsilon_i,$$

   where $w_k \sim N(0, \sigma_w^2)$.

Group effects are also called random effects:

1. Random effect for the intercept $u_k \sim N(0, \sigma_u^2)$
2. Random effect for the slope $w_k \sim N(0, \sigma_w^2)$

Random intercept

1. Random effect model with random intercept

$$y_i = (\alpha_0 + u_k) + \beta x_i + \epsilon_i, \ = \alpha_0 + \beta x_i + u_k + \epsilon_i,$$

where α_0 is the intercept and β the regression coefficient.

- There are two distinct error terms
  1. Group-specific error $u_k \sim N(0, \sigma_u^2)$
  2. Individual-specific error $\epsilon_i \sim N(0, \sigma^2)$
- Note that u_k and ϵ_i are independent of each other.

Random effect model with random intercept

Interpretation of random intercept α_k:

$$\alpha_k = (\alpha_0 + u_k)$$

- α_0 is the global intercept
- u_k group-level variations around the global intercept

This is equivalent to assuming α_k is a **random variable** that follows a Normal distribution

$$\alpha_k \sim N(\alpha_0, \sigma_u^2)$$

Random effect model with random intercept

Multi-level interpretation (two levels of variability):

1. **First level**. Defined on the individual level for observation $i = 1, ..., n$, similar to a standard linear regression
$$y_i = \alpha_k + \beta x_i + \epsilon_i$$
2. **Second level**. But the intercept is not fixed, it is a random variable
$$\alpha_k \sim N(\alpha_0, \sigma_u^2)$$

Random effect model with random intercept

Assumptions:

- Slope of regression line is the same across all groups. Each group has a different intercept (α_k).
- But $\alpha_k \sim N(\alpha_0, \sigma_u^2)$ has now a common distribution which is estimated from **all observations**, and not just from the observations in a specific group as in fixed effects.
- We pool information across groups.

Consequences:

- We control for group characteristics by including the group-specific intercept.
- Number of group-specific parameters to estimate is much smaller than in the fixed effect models (σ_u^2 vs k intercepts).

(Restricted) Maximum Likelihood estimation of random effect

$$y_i = \alpha_0 + \beta x_i + u_k + \epsilon_i$$

Parameters to estimate are

- α_0, β intercept and regression coefficient
- σ_u^2, σ^2 variance components

Maximum Likelihood estimation is based on the Normal distribution of u_k and ϵ_i

- ML estimate for σ_u^2 requires subtracting 2 empirical estimates of variance $\rightarrow$ ML estimates for σ_u^2 can be negative.
- Restricted Maximum Likelihood (REML): Imposes positivity constraints on the variance estimates.

Random effects in `R`: `nlme::lme()`

Implementations of Restricted Maximum Likelihood (REML) in `R`:

- `lmer` function in the `lme4` package
- `lme` function in the `nlme` package

Focus here is on `nlme::lme(fixed, random, data)`:

- `fixed`. Formula $y \sim x$
- `random`. Formula $\sim 1 \mid$ factor
- `data`. Dataset to use

`R`: Random intercept using `lme()`

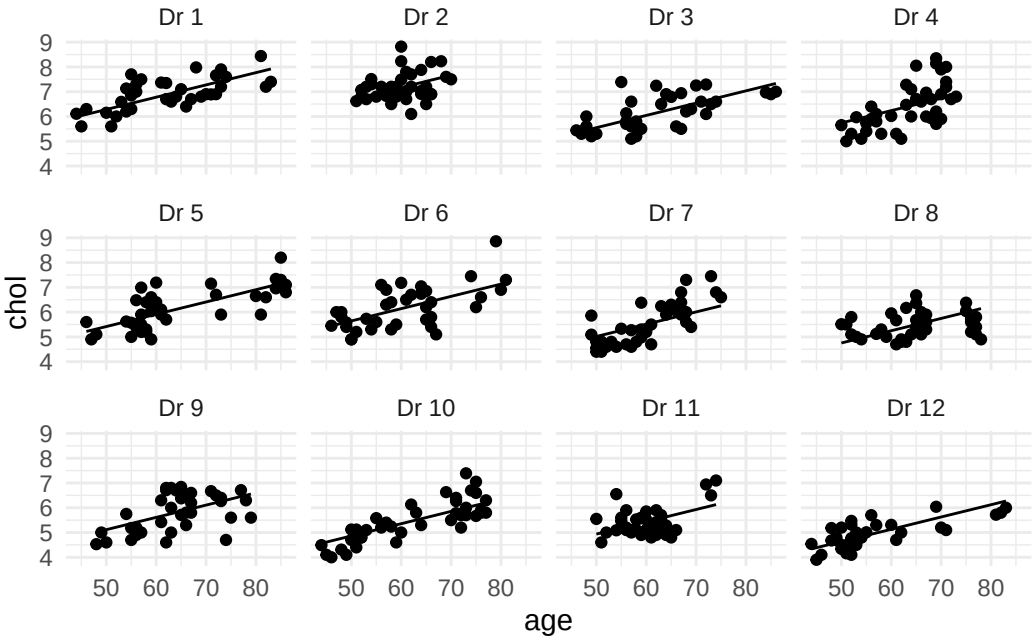
```
model_random_intercept <- lme(chol ~ age, random = ~ 1 | doctor, data = data_chol)
summary(model_random_intercept)
```

```
Linear mixed-effects model fit by REML
  Data: data_chol
      AIC     BIC    logLik
  828.697 845.035 -410.3485

Random effects:
 Formula: ~1 | doctor
        (Intercept)  Residual
```

```
StdDev:   0.6347908 0.5764246

Fixed effects:  chol ~ age
                Value  Std.Error  DF  t-value p-value
(Intercept) 2.9060357 0.26477408 428 10.97553       0
age         0.0495831 0.00306279 428 16.18888       0
 Correlation:
    (Intr)
age -0.714

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.7850636 -0.7012544 -0.1419474  0.6536599  3.0850909

Number of Observations: 441
Number of Groups: 12
```

R: Random intercept using `lme()`

```r
data_chol |>
  mutate(
    doctor_name = factor(str_c("Dr ", doctor), levels = str_c("Dr ", 1:12)),
    .fitted = fitted(model_random_intercept)
  ) |>
  ggplot(aes(x = age)) +
  geom_point(aes(y = chol)) +
  geom_line(aes(y = .fitted)) +
  facet_wrap(~doctor_name) +
  theme_minimal()
```



Random effect model and variance partition

Variance decomposition for observation i in group k

$$\begin{aligned} \text{var}(y_i) &= \text{var}(u_k + \epsilon_i) && (1)\\ &= \text{var}(u_k) + \text{var}(\epsilon_i) + 2\text{cov}(u_k, \epsilon_i) && (2)\\ &= \sigma_u^2 + \sigma^2 + 0 && (3) \end{aligned}$$

Further we can look at the covariance of observations

- i and i' within group k

$$\text{cov}(y_i, y_{i'}) = \text{cov}(u_k + \epsilon_i, u_k + \epsilon_{i'}) = \sigma_u^2$$

- i and i' from different groups k and k'

$$\text{cov}(y_i, y_{i'}) = \text{cov}(u_k + \epsilon_i, u_{k'} + \epsilon_{i'}) = 0$$

Random effect model and variance partition

Variability between and within groups. Intra-class correlation coefficient ρ

$$\rho = \text{cor}(y_i, y_{i'}) = \frac{\text{cov}(y_i, y_{i'})}{\sqrt{\text{var}(y_i)\text{var}(y_{i'})}} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

Interpretation:

- Intra-class correlation coefficient ρ is the correlation between two observations i and i' in the same group.
- It is the ratio of between-group variance σ_u^2 over the total variance.
- If $\rho \to 0$ there is little variation explained by the grouping and we might consider a model without the random effect.
- Any restrictions?

Variance partition in `R`

```
VarCorr(model_random_intercept)
```

```
doctor = pdLogChol(1)
            Variance  StdDev
(Intercept) 0.4029594 0.6347908
Residual    0.3322653 0.5764246
```

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} = \frac{0.6347908^2}{0.6347908^2 + 0.5764246^2} \approx 0.54$$

Interpretation:

- There is substantial evidence for between-group heterogeneity.
- More than half of the total variance can be explained by the between-group variance.
- It is beneficial to include the random effects on the intercept.

Random effect model with random intercept and random slope

Random effects model on both the intercept and the slope

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \epsilon_i$$

- There are three distinct error terms
  1. Group-specific error of the intercept
  $$u_k \sim N(0, \sigma_u^2)$$
  2. Group-specific error of the regression slope
  $$w_k \sim N(0, \sigma_w^2)$$
  3. Individual-specific error
  $$\epsilon_i \sim N(0, \sigma^2)$$
- Note that u_k and w_k are correlated and independent of ϵ_i.

Random effect model with random intercept and random slope

Assumptions:

- Each group has a different intercept ($\alpha_k = \alpha_0 + u_k$) and a different regression slope ($\beta_k = \beta + w_k$).
- We allow for correlation between α_k and β_k.
- Both, $\alpha_k \sim N(\alpha_0, \sigma_u^2)$ and $\beta_k \sim N(\beta, \sigma_w^2)$ have a common distribution which is estimated from **all observations**, and not just from the observations in a given group as in fixed effects.
- We pool information across groups.

Consequences:

- Including a random slope can be interpreted as creating an interaction between the group and the strength of association.
- We only have three additional parameters in the model: σ_u^2, σ_w^2 and $cor(\sigma_u, \sigma_w)$.

R: Random intercept and slope using `lme()`

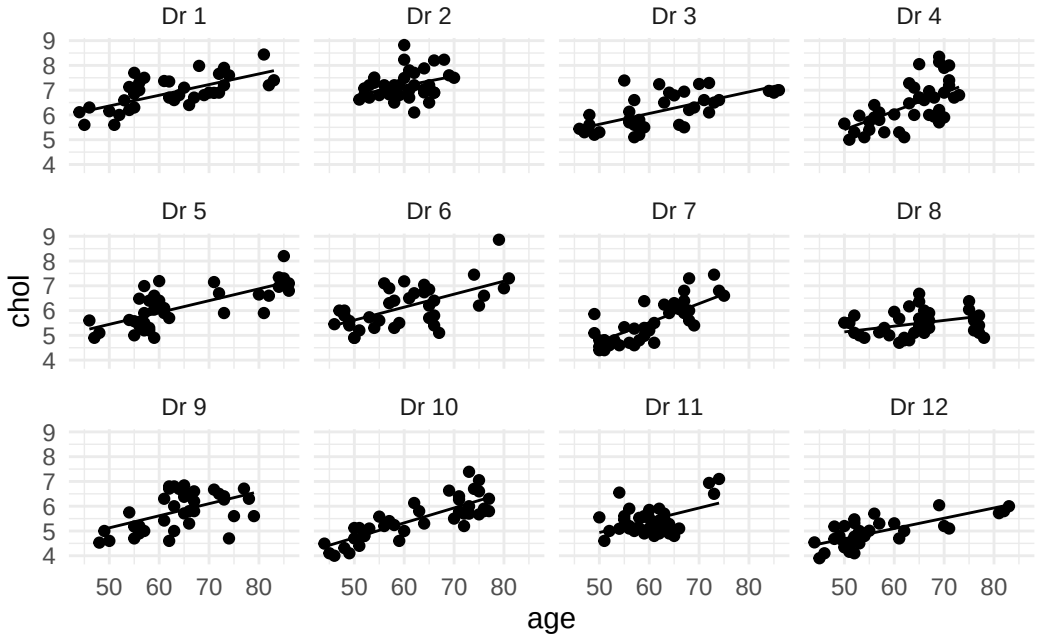
```r
model_random_slope <- lme(chol ~ age, random = ~ 1 + age | doctor, data = data_chol)
summary(model_random_slope)
```

```
Linear mixed-effects model fit by REML
  Data: data_chol
       AIC      BIC    logLik
  821.9886 846.4956 -404.9943

Random effects:
 Formula: ~1 + age | doctor
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 1.28163949 (Intr)
age         0.01771589 -0.872
Residual    0.55997507

Fixed effects:  chol ~ age
                Value Std.Error  DF  t-value p-value
(Intercept) 2.8791744 0.4215204 428 6.830451       0
age         0.0500704 0.0060597 428 8.262823       0
 Correlation:
    (Intr)
age -0.901

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.8523390 -0.6664557 -0.1141924  0.6206844  3.0809629

Number of Observations: 441
Number of Groups: 12
```

R: Random intercept and slope using `lme()`

```r
data_chol |>
  mutate(
    doctor_name = factor(str_c("Dr ", doctor), levels = str_c("Dr ", 1:12)),
    .fitted = fitted(model_random_slope)
  ) |>
```

```
ggplot(aes(x = age)) +
geom_point(aes(y = chol)) +
geom_line(aes(y = .fitted)) +
facet_wrap(~doctor_name) +
theme_minimal()
```



Variables on individual and group level

When considering variables or predictors we need to distinguish:

- **Individual-level** variables
- **Group-level variables** that are the same for all observations in a group

GP example:

- Individual-level variables: Age and sex of patient
- Group-level variables: Age of doctor

```
# A tibble: 6 x 6
   chol doctor   age   bmi agedoc   sex
  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl>
1  7.13      1    54  27.4     55     0
```

```
2  7.7       1    55  29.1     55     0
3  7.3       1    56  27.9     55     0
4  6.89      1    71  26.7     55     1
5  6.9       1    72  26.7     55     1
6  7.9       1    73  29.7     55     1
```

Variables on individual and group level

$$y_i = (\alpha_0 + u_k) + (\beta + w_k)x_i + \gamma x_g + \epsilon_i$$

```
model_random_cov <- lme(chol ~ age + agedoc, random = ~ 1 + age | doctor, data = data_chol
summary(model_random_cov)
```

```
Linear mixed-effects model fit by REML
  Data: data_chol
       AIC      BIC    logLik
  815.6956 844.2712 -400.8478

Random effects:
 Formula: ~1 + age | doctor
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev     Corr
(Intercept) 1.10559524 (Intr)
age         0.01775586 -0.951
Residual    0.55992053

Fixed effects:  chol ~ age + agedoc
                 Value Std.Error  DF   t-value p-value
(Intercept) -2.7897788 1.1824050 428 -2.359410  0.0188
age          0.0501492 0.0060673 428  8.265423  0.0000
agedoc       0.1280030 0.0253576  10  5.047908  0.0005
 Correlation:
       (Intr) age
age    -0.303
agedoc -0.948 -0.004

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.8426263 -0.6665146 -0.1046213  0.6206861  3.0950474

Number of Observations: 441
```

```
Number of Groups: 12
```

Model comparison

Likelihood-ratio test for nested models:

- Models must have the same fixed effects. Does not work with group-level covariates.
- Model with smaller - log likelihood is better (better model fit).

Akaike information criterion (AIC):

- Model with the smaller AIC is better (less information loss).

Model comparison

GP example:

- Model A (Random intercept) `model_random_intercept = lme(chol ~ age, random = ~ 1 | doctor, data = data_chol)`
- Model B (Random intercept and slope) `model_random_slope = lme(chol ~ age, random = ~ 1 + age | doctor, data = data_chol)`
- Model C (Random intercept and slope and group covariate) \ `model_random_cov = lme(chol ~ age + agedoc, random = ~ 1 + age | doctor, data = data_chol)`

Model comparison

Likelihood-ratio test for nested models:

```
anova(model_random_intercept, model_random_slope)
```

```
                      Model df      AIC      BIC    logLik   Test L.Ratio
model_random_intercept     1  4 828.6970 845.0350 -410.3485
model_random_slope         2  6 821.9886 846.4956 -404.9943 1 vs 2 10.7084
                      p-value
model_random_intercept
model_random_slope     0.0047
```

Akaike information criterion (AIC):

```
anova(model_random_slope, model_random_cov)
```

```
Warning in anova.lme(model_random_slope, model_random_cov): fitted objects with
different fixed effects. REML comparisons are not meaningful.
```

```
                   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model_random_slope     1  6 821.9886 846.4956 -404.9943
model_random_cov       2  7 815.6956 844.2712 -400.8478 1 vs 2 8.292926   0.004
```

Generalised linear mixed models

- Generalised Linear Mixed models (GLMM) can be used to adapt linear mixed models to outcomes that do not follow a Normal distribution.
- The package `lme4` includes the function `glmer` that can fit GLMMs.

```
glmer(formula = y ~ x + (1 + x | factor), family = gaussian)
```

Take away: Fixed and random effects

- Fixed effect models can account for group structure but many parameters need to be estimated and no information is shared between groups.
- Random effect models treat group-specific parameters as random variables.
- Instead of estimating one parameter for each group, random effect models only estimate the distribution parameter of the random variable.
- Thus, they pool information across groups.
- The intra-class coefficient gives a measure of how relevant the group structure is.
- Implementation in `R`: `lme()` function in the `nlme` package.
- Models including both, fixed and random effects, are often called linear mixed models.